12-1999

On-Line Bayesian Tree-Structured Transformation of Hidden Markov Models for Speaker Adaptation

Shaojun Wang
Wright State University - Main Campus, shaojun.wang@wright.edu

Yunxin Zhao

ON-LINE BAYESIAN TREE-STRUCTURED TRANSFORMATION OF HIDDEN MARKOV MODELS FOR SPEAKER ADAPTATION

Shaojun Wang[1] *Yunxin Zhao*[2]

Beckman Institute and Dept. of ECE, University of Illinois, Urbana, IL 61801[1]
Dept. of CECS, University of Missouri, Columbia, MO 65211[2]
swang@ifp.uiuc.edu zhao@cecs.missouri.edu

ABSTRACT

This paper presents a new recursive Bayesian learning approach for transformation parameter estimation in speaker adaptation. Our goal is to incrementally transform (or adapt) the entire set of HMM parameters for a new speaker or new acoustic enviroment from a small amount of adaptation data. By establishing a clustering tree of HMM Gaussian mixture components, the finest affine transformation parameters for individual HMM Gaussian mixture components can be dynamically searched. The on-line Bayesian learning technique proposed in our recent work is used for recursive maximum a posteriori estimation of affine transformation parameters. Speaker adaptation experiments using a 26-letter English alphabet vocabulary are conducted, and the viability of the on-line learning framework is confirmed.

1. INTRODUCTION

Adaptation technique has been widely studied for practical speech recognition systems in the last decade. It can be classified into the following two major approaches: Bayesian approach [4] and transformation approach [2, 6]. In the Bayesian adaptation approach [4], prior distributions are assumed for the parameters in HMMs and maximum a posteriori (MAP) estimates for the parameters are calculated instead of maximum likelihood (ML) estimates. In this approach, recognition performance improvement will be achieved when the amount of adaptation data is relatively large. In transformation based adaptation approach, a simple transformation is defined in the spectral feature space or the HMM parameter space and the transformation parameters are estimated using the adaptation data. Transformation approach is more effective than Bayesian approach when the amount of adaptation data is relatively small.

Adaptation techniques may operate in a number of *modes*: *supervised* vs. *unsupervised* and *batch* vs. *on-line* (*incremental* or *sequential*). For real-world applications, the unsupervised mode is usually more realistic and desirable. The on-line adaptation is aiming at incrementally estimating the parameters as each block of adaptation/training/testing data is enrolled. This block data is then discarded after adaptation is completed. As a consequence, the computational load and memory requirement can be effectively reduced.

Modeling the correlations among speech sounds for speaker adaptation is widely studied in recent years. Since it is unlikely to have sufficient speech data corresponding to all HMM units in a small adaptation set, certain parameter *correlation* and *tying* are introduced so that the model parameters can be consistently and fully adjusted. This agrees in principle with the notion that there exists an underlying acoustic and linguistic structure in speech that should be exploited to improve adaptation efficiency. *Correlation* among Gaussian mean parameter vectors have been used in HMM parameter adaptation. *Tying* is widely used in transformation based adaptation [6]. Certain techniques relate parameter vectors across all classes by making Markovian assumptions on the dependency structure. The joint correlation is represented by a low-order conditional distribution and hence a relatively small number of parameters are used to characterize the dependency. Examples include Markov random fields, multiscale tree processes, tree-structural MAP adaptation.

In this paper, we propose an on-line Bayesian transformation algorithm that uses a hierarchical tree structure to control the degree of transformation tying. The underlying theoretical framework is the recursive Bayesian learning we developed recently. Chien [1] recently developed an on-line Bayesian transformation-based adaptation scheme, that uses the recursive approximate quasi-Bayes algorithm [5] to estimate the model transformation parameters. The HMM Gaussian parameters were incrementally adapted through a set of transformation functions. A hierarchical tree of HMM parameters was build, and for each HMM Gaussian component, the nearest node containing the adaptation data tokens was extracted and its parameters were used for on-line adaptation. Due to the use of recursive quasi-Bayes estimation, the forms of transformation functions were limited to those having a reproducible prior/posterior pdf pairs, which unfortunately were very few. The recognition performance was sensitive to the update intervals, the longer the better. Using long update intervals is usually undesirable due to the associated high computational cost and expensive memory storage. Our proposed on-line Bayesian learning overcomes these drawbacks.

2. RECURSIVE BAYESIAN LEARNING

Assume that independent random variables $\underline{o}_k$'s are receiveded sequentially and are described by a probability density function with parameter η. Applying the Bayes theorem, we obtain a recursive expression for the a posteriori pdf of η, given $\underline{o}^k = \{\underline{o}_1, \cdots, \underline{o}_k\}$, as

$$p(\eta|\underline{o}^k) = \frac{p(\underline{o}_k|\underline{o}^{k-1},\eta)p(\eta|\underline{o}^{k-1})}{p(\underline{o}_k|\underline{o}^{k-1})} = \frac{p(\underline{o}_k|\eta)p(\eta|\underline{o}^{k-1})}{\int p(\underline{o}_k|\eta)p(\eta|\underline{o}^{k-1})d\eta} \quad (1)$$

Successive computation of Eqn. (1) introduces an ever-expanding combination of the previously obtained posterior pdf's and thus quickly leads to a combinatorial explosion of product terms. Quasi-Bayes learning [5] approximates the resulting posterior distribution $p(\eta|\underline{o}^k)$, by the "closest" tractable distribution $g(\eta|\phi^{(k)})$ with a given class P, under the criterion that both distributions have the same mode. Then the EM algorithm is applied and the hyper-parameters of the approximate posterior distribution and model parameters are incrementally updated.

In this work, we propose a new approach for recursive Bayesian learning of parameters of stochastic processes. At each step, we only calculate η that maximizes $p(\eta|\underline{o}^k)$. Repeated maximization of $p(\eta|\underline{o}^k), k = 1, 2, \cdots$, produces the sequence of estimates.

The general recursive estimator for parameters of a stochastic process is of the form:

$$\eta^{(k+1)} = \eta^{(k)} + \varepsilon_k H(\underline{o}^{k+1}, \eta^{(k)})^{-1} h(\underline{o}_{k+1}, \eta^{(k)}) \quad (2)$$

where ε_k is a sequence of small positive gains which can be either fixed as a constant or decrease with the index k. The adaptive matrix $H(\underline{o}^{k+1}, \eta^{(k)})$ and the score function

$h(\underline{o}_{k+1}, \eta^{(k)})$ jointly define the way that the parameter η is updated as a function of new observations.

Define $R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ to be the auxiliary function of log posterior likelihood, $R_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) = Q_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) + log p(\eta|\phi)$, where $Q_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ denotes the auxiliary function of log likelihood as defined in EM algorithm [7] and $p(\eta|\phi)$ is the prior pdf of η with parameter ϕ. We first derive a recursive estimation formula by using the normalized auxiliary function $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ as the objective function. Maximizing the second-order Taylor series expansion of $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ with respect to η and denoting the maximizing point by $\eta^{(k+1)}$, we have

$$\eta^{(k+1)} = \eta^{(k)} + [H(\underline{o}^{k+1}, \eta^{(k)})]^{-1} \frac{1}{k+1} \frac{\partial R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}} \quad (3)$$

and

$$\begin{aligned} H(\underline{o}^{k+1}, \eta^{(k)}) &= -\frac{1}{k+1} \frac{\partial^2 R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})}{\partial \eta \partial \eta^T} \\ &= \frac{1}{k+1} [I_C(\underline{o}^{k+1}|\eta^{(k)}) + I_p(\eta^{(k)})] \end{aligned} \quad (4)$$

where $I_C(\underline{o}^{k+1}|\eta^{(k)}) = \sum_{i=1}^{k+1} I_C(\underline{o}_i|\eta^{(k)})$ is the conditional expectation of the complete-data information matrix given $\underline{o}^{k+1}$ [7], $I_p(\eta)$ is prior information matrix, i.e., negative Hessian matrix of log $p(\eta|\phi)$. Eqn. (3) corresponds to one step of the Newton-Raphson algorithm for maximizing the normalized auxiliary function $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$, initialized at $\eta^{(k)}$, and it is called the (batch) EM gradient algorithm [7]. From Eqn. (4), we can see that the effect of prior information decreases as the number of observations becomes large. If we replace $I_C(\underline{o}^{k+1}|\eta^{(k)})$ in $H(\underline{o}^{k+1}, \eta^{(k)})$ by its expectation, i.e., the complete data Fisher information $I_{CF}^{k+1}(\eta^{(k)}) = \sum_{i=1}^{k+1} I_{CF,i}(\eta^{(k)})$, then Eqn. (3) corresponds to one step of the (batch) EM gradient scoring algorithm [7].

Let $\ell_k(\eta, \eta^{(k)}) = R_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) - R_{\underline{o}^k}(\eta, \eta^{(k)}) = Q_{\underline{o}_{k+1}}(\eta, \eta^{(k)})$. Then, $\frac{\partial R_{k+1}(\eta, \eta^{(k)})}{\partial \eta} = \frac{\partial R_k(\eta, \eta^{(k)})}{\partial \eta} + \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}$. Assuming that $\eta^{(k)}$ maximizes $R_k(\eta, \eta^{(k)})$, so that $\frac{\partial R_{\underline{o}^k}(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}} = 0$, then we have

$$\frac{\partial R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}} = \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}} \quad (5)$$

As such, we obtain a recursive estimation formula as the following:

$$\eta^{(k+1)} = \eta^{(k)} + [H(\underline{o}^{k+1}, \eta^{(k)})]^{-1} \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}} \quad (6)$$

where $H(\underline{o}^{k+1}, \eta^{(k)})$ is recursively calculated by an approximation $\frac{1}{k+1}[\sum_{i=1}^{k+1} I_C(\underline{o}_i|\eta^{(i)}) + I_p(\eta^{(k)})]$ or $\frac{1}{k+1}[\sum_{i=1}^{k+1} I_{CF,i}(\eta^{(i)}) + I_p(\eta^{(k)})]$. In order to satisfy $\frac{\partial R_{\underline{o}^{k+1}}(\eta, \eta^{(k+1)})}{\partial \eta}|_{\eta=\eta^{(k+1)}} = 0$, we iterate Eqn. (6) several times for each $\eta^{(k)}$ and denote the resulting estimate as $\eta^{(k+1)}$.

When the complete data belongs to regular exponential family, the information matrix $I_C(\underline{o}^{k+1}|\eta^{(k)})$ or the complete data Fisher information $I_{CF}^{k+1}(\eta^{(k)})$ is indeed always positive definite. If we choose the prior probability properly, the matrix $H(\underline{o}^{k+1}, \eta^{(k)})$ will be positive definite. This fact in turn implies strict concavity of $R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ and uniqueness of the maximizing point for $R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$. It also implies that $H(\underline{o}^{k+1}, \eta^{(k)})^{-1} \frac{1}{k+1} \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}|_{\eta=\eta^{(k)}}$ is an ascent direction, and therefore a fractional step ε in the direction will lead to an increase in $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\underline{o}^{k+1}, \eta)$, that is, $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\underline{o}^{k+1}, \eta^{(k+1)}) > \frac{1}{k+1} R_{\underline{o}^{k+1}}(\underline{o}^{k+1}, \eta^{(k)})$. Thus we can modify Eqn. (6) to be

$$\eta^{(k+1)} = \eta^{(k)} + \varepsilon_k H(\underline{o}^{k+1}, \eta^{(k)})^{-1} \frac{1}{k+1} \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}]|_{\eta=\eta^{(k)}} \quad (7)$$

and the modified algorithm has the desirable property of being locally monotonic when $0 < \varepsilon_k < 2$. The optimal choice of ε_k at each step is determined by line search along the direction of $H(\underline{o}^{k+1}, \eta^{(k)})^{-1} \frac{1}{k+1} \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}]|_{\eta=\eta^{(k)}}$ from the current estimate $\eta^{(k)}$ to maximize $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$.

3. RECURSIVE BAYESIAN LEARNING FOR TRANSFORMATION PARAMETERS

Among the previous studies, Digalakis [3] applied the incremental EM algorithms in recursive maximum likelihood estimation of affine transformation parameters. Huo and Lee developed a quasi-Bayes approach [5] for HMMs. Chien [1] applied the quasi-Bayes approach for on-line tree-structured transformation of HMM. In Chien's work, the form of transformation function was restricted in order to derive a reproducible prior/posterior pair. In our current study, the recursive Bayesian learning is considered for on-line estimation of transformation parameters. Due to the relaxation in the form of priors, the transformation function can be chosen in many forms without complicating the algorithm.

Consider an N-state continuous density HMM. The state observation probability density function of an vector observation $Y_t \in \Re^n$ at time t is assumed to be a mixture of M multivariate Gaussian distributions:

$$p(y_t|x_t = i; \lambda_i) = \sum_{m=1}^{M} c_{i,m} N(y_t|\mu_{i,m}, \Sigma_{i,m}) \quad (8)$$

where $N(y_t|\mu_{i,m}, \Sigma_{i,m})$ denotes a Gaussian source with mean vector $\mu_{i,m}$ and covariance matrix $\Sigma_{i,m}$, $c_{i,m}$ denotes mixture weight, and the parameters of the mixture densities are denoted by $\lambda = \{\lambda_{i,m}\} = \{c_{i,m}, \mu_{i,m}, \Sigma_{i,m}\}, i = 1, \cdots, N, m = 1, \cdots, M$.

In recursive transformation-based Bayesian learning, either the HMM parameters λ are transformed according to a transformation function $G_\eta(\cdot)$ in the model space, or the observations are obtained through a transformation function $F_\eta(\cdot)$ in the feature space, where $\eta \in \Re^D$ represents *nuisance* parameters to be estimated. Let $\underline{o}^k = \{\underline{o}_1, \cdots, \underline{o}_k\}$ be k successively independent blocks of speech feature vectors. The data blocks are used as the adaptation data to estimate the transformation parameters η.

Several transformation functions have been introduced in literature for compensating the mismatch between test speech and trained model speech. Extensive studies show that affine transformation gives rather large improvement to speech recognition performance [2]. However, previous efforts in on-line Bayesian estimation could only consider bias and variance transformations due to the constraint in reproducible prior/posterior pairs. In the current study, affine transformation is employed, and the transformation of HMM parameters $\lambda_{i,m} = (c_{i,m}, \mu_{i,m}, \Sigma_{i,m})$ has the form of

$$\hat{\lambda} = G_{\eta^{(k)}} = \left(c_{i,m}, A_c^{(k)} \mu_{i,m} + b_c^{(k)}, A_c^{(k)} \Sigma_{i,m} (A_c^{(k)})^T\right) \quad (9)$$

Transformation function $G_\eta(\cdot)$ have C clusters with $\eta^{(k)} = \{(A_c^{(k)})^{-1}, b_c^{(k)}\}, c = 1, \cdots, C$, and the HMM parameters $\lambda_{i,m} = (\mu_{i,m}, \Sigma_{i,m})$ are assumed to be labeled by the cth cluster membership Ω_c.

In choosing the prior pdf for η, we assume that the matrix $\hat{A}_c$, the inverse of transformation matrix A_c, and the bias vector b_c are jointly Gaussian, i.e., $(\hat{A}_c, b_c) \sim N(\hat{A}_c, b_c|\phi_c = \{\mu_{\hat{A}_c, b_c}, \Sigma_{\hat{A}_c, b_c}\})$.

In this work we use the EM gradient recursive Bayesian learning for transformation parameters. Assume that the length of the $k+1$th data block is T_{k+1}, i.e., $\underline{o}_{k+1} = (o_{k+1,1}, \cdots, o_{k+1,T_{k+1}})$. Then

$$\ell_k(\eta, \eta^{(k)}) = Q_{\underline{o}_{k+1},c}(\eta, \eta^{(k)}) = -\frac{1}{2}\sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\xi_t^k(i,m)\Big(log|\Sigma_{i,m}| - 2log|\hat{A}_c^{(k)}| + (\hat{A}_c^{(k)}o_{k+1,t} - \mu_{i,m} - \hat{A}_c^{(k)}b_c^{(k)})^T \Sigma_{i,m}^{-1}(\hat{A}_c^{(k)}o_{k+1,t} - \mu_{i,m} - \hat{A}_c^{(k)}b_c^{(k)})\Big) \tag{10}$$

where $\xi_t^k(i,m) = Pr(x_t^{(k)} = i, z_t^{(k)} = m|\underline{o}_{k+1}, A_c^{(k)}, b_c^{(k)})$.

Thus the score statistic can be written as

$$\frac{\partial \ell_k(\eta,\eta^{(k)})}{\partial \hat{A}_c^{(k)}} = -\sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\xi_t^k(i,m)\Big(-[(\hat{A}_c^{(k)})^{-1}]^T + \Sigma_{i,m}^{-1}(\hat{A}_c^{(k)}o_{k+1,t} - \mu_{i,m} - \hat{A}_c^{(k)}b_c^{(k)})(o_{k+1,t} - b_c^{(k)})^T\Big) \tag{11}$$

$$\frac{\partial \ell_k(\eta,\eta^{(k)})}{\partial b_c^{(k)}} = \sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\xi_t^k(i,m)\Big((\hat{A}_c^{(k)})^T\Sigma_{i,m}^{-1}(\hat{A}_c^{(k)}o_{k+1,t} - \mu_{i,m} - \hat{A}_c^{(k)}b_c^{(k)})\Big) \tag{12}$$

Define $\psi_{t,i}^{(k)} = p(x_t = i|\lambda^{(k)})$, which can be obtained by transition matrix multiplication $(\pi\tilde{A}(\lambda^{(k)})^t)_i$, $r_{i,m} = \Sigma_{i,m}^{-1}$, i.e., the precision matrix, with $r_{i,m,j,l}$, the (j,l)th element of $r_{i,m}$, then the Fisher information matrix $I_{CF,k+1,c}(\eta^{(k)})$ can be obtained by $-E[\partial^2 logp(\underline{o}_{k+1}, \underline{x}_{k+1}, \underline{z}_{k+1})/\partial\eta\partial\eta^T]|_{\eta=\eta^{(k)}}$, where $\underline{x}_{k+1}$ and $\underline{z}_{k+1}$ denote the state and mixture indices for $k+1$th block of speech feature vectors, respectively.

$$I_{C,\underline{o}_{k+1},c}(\hat{a}_{c,j,l}, \hat{a}_{c,p,q}) = \sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\psi_{t,i}^{(k)}c_{i,m}^{(k)} (-a_{c,l,p}a_{c,q,j} + 2[r_{i,m,j,p}\sum_{w=1}^{n}\sum_{s=1}^{n}\hat{a}_{c,l,s}r_{i,m,s,w}\hat{a}_{c,q,w}]\Big) \tag{13}$$

for $j,l,p,q = 1,\cdots,n$

$$I_{C,\underline{o}_{k+1},c}(\hat{a}_{c,j,l}, b_{c,p}) = \sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\psi_{t,i}^{(k)}c_{i,m}^{(k)} \Big(-2\sum_{s=1}^{n}\hat{a}_{c,s,p}r_{i,m,j,s}\sum_{w=1}^{n}\hat{a}_{c,l,w}\mu_{i,m,w}\Big) \tag{14}$$

for $j,l,p = 1,\cdots,n, l\neq p$.

$$I_{C,\underline{o}_{k+1},c}(\hat{a}_{c,j,l}, b_{c,p}) = \sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\psi_{t,i}^{(k)}c_{i,m}^{(k)} \Big(-2\sum_{s=1}^{n}r_{i,m,j,s}[(\hat{a}_{c,s,l}\sum_{w=1}^{n}a_{c,l,w} + \sum_{v=1}^{n}\hat{a}_{c,s,v}\hat{a}_{c,v,w})\mu_{i,m,w} - \mu_{i,m,s}]\Big) \tag{15}$$

for $j,l = p = 1,\cdots,n$.

$$I_{C,\underline{o}_{k+1},c}(b_{c,p}, b_{c,q}) = \sum_{t=1}^{T_{k+1}}\sum_{i,m\in\Omega_c}\psi_{t,i}^{(k)}c_{i,m}^{(k)} \Big(2\sum_{w=1}^{n}\sum_{s=1}^{n}\hat{a}_{c,s,p}r_{i,m,s,w}\hat{a}_{c,w,q}\Big) \tag{16}$$

for $p,q = 1,\cdots,n$

Of course, we can also use the matrix $I_C(\underline{o}_{k+1}|\eta^{(k)})$, negative of the second derivative of Eqn. (10).

The prior information matrix $I_c^p(\eta)$ is given as

$$I_c^p(\eta) = -\partial^2 logp(\eta)/\partial\eta\partial\eta^T = r_{\hat{A}_c,b_c} \tag{17}$$

which is the precision matrix for $(\hat{A}_c, b_c)$, i.e., $r_{\hat{A}_c,b_c} = \Sigma_{\hat{A}_c,b_c}^{-1}$.

Using $\eta^{(k+1)} = \{(A_c^{(k+1)})^{-1}, b_c^{(k+1)}\}, c = 1,\cdots,C$, the HMM parameters are transformed according to Eqn. (9).

4. PRIOR DENSITY ESTIMATION

Generally the empirical Bayes approach is used to estimate the parameters ϕ_c of the prior density [4]. Let $O_1,\cdots,O_S$ denote the training data of speakers $s = 1,\cdots,S$, each has the transformation parameter $\eta_{c,s}$ and the $\eta_{c,s}$' have a common prior distribution $p(\eta_c|\phi_c)$. The marginal distribution of training data O can be written as

$$p(O|\phi_c) = \int\prod_{s=1}^{S}p(O_s|\eta_{c,s})p(\eta_{c,s}|\phi_c)d\eta_{c,s} \tag{18}$$

Based on $p(O|\phi_c)$, we obtain a maximum likelihood estimate $\hat{\phi}_c$. However, $\hat{\phi}_c$ is rather difficult to obtain due to the integration in Eqn. (18). To alleviate the problem, we use an alternative maximization procedure over η_c and ϕ_c as suggested in [4], i.e.

$$\eta_c^{(k)} = argmax_{(\eta_c)}p(O,\eta_c|\phi_c^{(k)}) \tag{19}$$

$$\phi_c^{(k+1)} = argmax_{(\phi_c)}p(\eta_c^{(k)}|\phi_c) \tag{20}$$

where Eqn. (19) is solved by the batch EM gradient method of Eqn. (3), and Eqn. (20) is straightforward to solve.

5. TREE-STRUCTURED TRANSFORMATION

When affine transformation functions are tied across Gaussian mixture components, each transformation function is associated with a number of mixture components. This is achieved by defining a set of transformation clusters where each cluster contains the mixture components associated with the same transformation function. The tree-structured clustering technique provides a hierarchical way in defining transformation clusters. Once a hierarchical tree is established, we are able to search for the most fitted transformation parameters for on-line transformation of HMM parameters and dynamically control the number of transformation parameters such that the recognition performance can be improved for limited adaptation data as well as abundant adaptation data.

In this work, we follow the same procedure as in Chien [1] where the Gaussian mixture components of HMM's are clustered by using the binary split K-means algorithm with divergence measure. The number of clusters C is chosen to be $2^d - 1$. After the clustering, a hierarchical tree of Gaussian parameters with d layers is built. Each Gaussian mixture component parameters $\lambda_{i,m}$ corresponds to d nodes in the tree, one at each layer. The root node covers all Gaussian mixture components and the leaf nodes are occupied by individual Gaussian mixture components.

In the hierarchical tree, the labels of Gaussian mixture components in each layer are stored in the corresponding tree node. Based on the labels, we can calculate the transformation parameters of each tree node by using the proposed on-line Bayesian learning. In general, the parameters in higher layers serve as *coarse transformation* and those in lower layers serve as *fine transformation*. To retain fine details of acoustic models, the transformation functions should be as detail as possible. In [1], a *bottom-up* strategy is proposed to automatically search for the transformation parameters of each Gaussian mixture component. In contrast, we use a white-black tree based *bottom-up top-down* strategy, where a *bottom-up* procedure is first used to perform on-line Bayesian learning of transformation parameters $\eta_c^{(k)}$ for the nodes containing adaptation data, and a *top-down* procedure is next used to perform affine transformations on all the HMM's Gaussian mixture components.

In the *bottom-up* procedure, we first mark all the nodes by white. We then find all the Gaussian mixture components with corresponding adaptation data and mark their parents by black. We then move up one layer and for each black node in this layer, we collect adaptation data and perform on-line Bayesian learning of transformation parameters $\eta_c^{(k)}$, and mark the parent of the node by black. We repeat this procedure until root node.

In the *top-down* procedure, starting from the root node, we go down one layer and check for white nodes from left to right. If the parent of the current white node is black, it indicates that the parent node has the *closest* (or *finest*) transformation parameters for the leaves of the white node, and thus we perform affine transformation on all the Gaussian mixture components covered under this node by using the transformation parameters at this parent node. If the parent of the current white node is white, we move to the next node at this layer. We repeat this procedure for all d layers. Figure 1 gives an illustration for this bottom-up estimation and top-down transformation procedure.

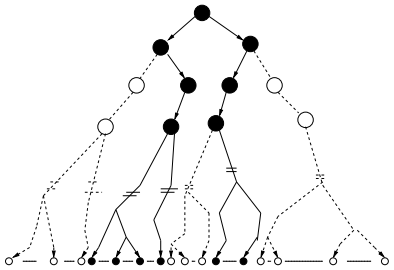

Figure 1. White-Black Tree for Bottom-Up Estimation and Top-Down Transformation

After this *bottom-up-top-down* procedure, all the Gaussian mixture components are updated by the *finest* transformation parameters available.

6. EXPERIMENTAL RESULTS

To examine the viability of the proposed technique, the on-line Bayesian learning approach is applied to on-line speaker adaptation. We report a series of recognition experiments using vocabulary of 26-letter English alphabet. Two severely mismatched speech databases , the OGI ISOLET and the TI46, were used for evaluating the adaptation algorithm. A full discription for these two corpora can be found in [5]. For speaker independent training, the OGI ISOLET data base was used. It consists of 150 speakers, each speaking each of the letters twice. For on-line Bayesian adaptation and testing, the English alphabet subset of the TI46 isolated word corpus was used. It was produced by 16 speakers (8 males and 8 females), among them, data from 4 males were incomplete. Therefore, only 12 speakers were used in this study. Each person uttered each of the letters 26 times. Ten of them were collected in the same session. The remaining 16 tokens were collected in 8 different sessions in which two tokens of each letter were collected in each session. For each person and each letter, we divide equally those 16 tokens collected in eight different sessions into two parts, one for adaptive training, another for testing.

In all the experiments, each letter in the vocabulary was modeled by a single left-to-right five-state CDHMM. Each state had a Gaussian mixture density of four components with each component density having a diagonal covariance matrix. The speech feature was extracted based on a tenth-order LPC analysis, where the feature components are 12 cepstral coefficients, a normalized log energy and their first time derivatives.

Starting with the SI initial models, we select training tokens for each letter randomly and perform utterance-based supervised on-line adaptation. After each adaptation, we test the recognizer on a seperate testing set to measure the performance changes. We plot in Fig. 2 the performance comparison of two setups, averaged over 12 speakers, as a function of average total number of adaptation tokens per speaker. In these two setups, one is using block diagonal transformation matrix (two blocks, one for cepstral coefficients and energy, the other is for the derivatives), another is using diagonal transformation matrix. In order to show the advantage of our approach, we also draw the results of

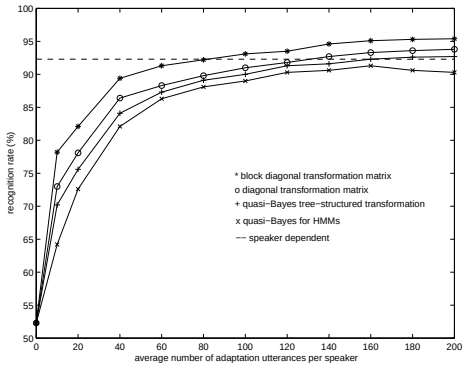

Figure 2. Performance of Recognition Results.

quasi-Bayes on-line estimation for tree-structured transformation and quasi-Bayes on-line estimation for HMMs from [5]. The recognition result is proven better than those of quasi-Bayes approaches. The main reason is that our recursive Bayesian learning technique accommodates easily hierarchical-tree based affine transformations and hence is effective for both limited adaptation data as well as for abundant adaptation data. Since block diagonal transformation matrix can consider rotations among elements of the feature vector, it gives better performance than that of diagonal transformation matrix.

7. CONCLUSION

In this work, we developed an on-line Bayesian learning technique for transformation of parameters of Gaussian densities of hidden Markov models. A hierarchical tree of HMM Gaussian parameters is employed to dynamically control the transformation tying. This technique updates parameter estimates after each utterance and can accommodate flexible forms of transformation functions as well as prior probability density functions. In the speaker adaptation evaluation, recognition accuracy was asymptotically improved for increasing number of adaptation data, and the performance was significantly superior to other on-line adaptation methods.

8. ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under grant NSF IRI-95-02074.

REFERENCES

[1] J. Chien, "On-Line Hierarchical Transformation of Hidden Markov Models for Speaker Adaptation," Proceedings of ICSLP98, pp. 2295-2298, 1998

[2] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," IEEE Trans. on Speech and Audio Processing, Vol. 3, pp. 357-366, September 1995

[3] V. Digalakis, "On-Line Adaptation of Hidden Markov Models Using Incremental Estimation Algorithms," IEEE Trans. on Speech and Audio Processing, pp. 253-261, May 1999

[4] J. Gauvain and C. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, April 1994

[5] Q. Huo and C. Lee, "On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate," IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 2, pp. 161-172, March 1997

[6] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, Vol. 9, pp. 171-185, 1995

[7] G. McLachlan and T. Krishnan, The EM Algorithm and Extensions, John Wiley & Sons, New York, 1997